(19)日本国特許庁（ＪＰ）　(12)公開特許公報（Ａ）　(11)特許出願公開番号

特開平10－247093

(43)公開日　平成10年(1998)９月14日

(51)Int.Cl.[6]		識別記号	FI		
Ｇ１０Ｌ	3/00	５３１	Ｇ１０Ｌ	3/00	５３１Ｎ
	7/02			7/02	Ａ
	9/00			9/00	Ｄ
// Ｈ０４Ｎ	7/24		Ｈ０４Ｎ	7/13	Ｚ

審査請求　未請求　請求項の数８　ＦＤ　（全９頁）

(21)出願番号　特願平9－65595

(22)出願日　平成９年(1997)３月５日

(71)出願人　000001214
国際電信電話株式会社
東京都新宿区西新宿２丁目３番２号

(72)発明者　氏原　清乃
東京都新宿区西新宿２丁目３番２号　国際電信電話株式会社内

(72)発明者　中島　康之
東京都新宿区西新宿２丁目３番２号　国際電信電話株式会社内

(72)発明者　米山　暁夫
東京都新宿区西新宿２丁目３番２号　国際電信電話株式会社内

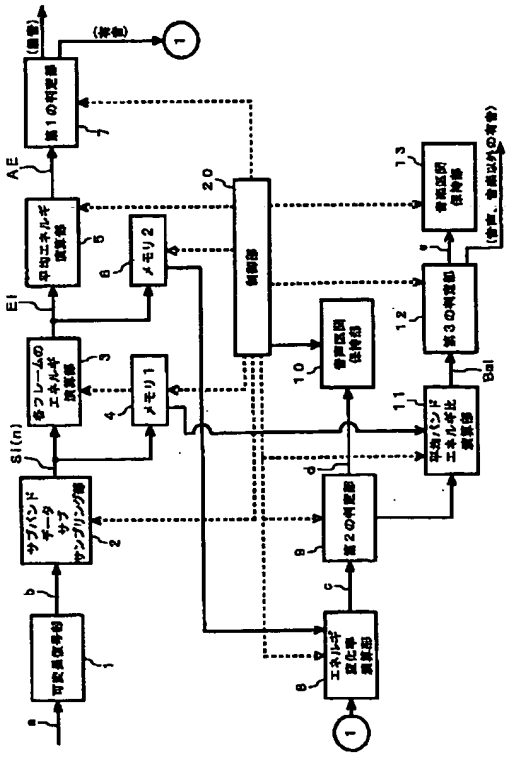
(74)代理人　弁理士　田中　香樹　（外１名）

(54)【発明の名称】　オーディオ情報分類装置

(57)【要約】

【課題】　音楽区間のみならず音声区間の検出もできるオーディオ情報分類装置を提供することにある。

【解決手段】　第１に、入力されたオーディオ情報から１秒毎の周波数データを抽出する。第２に、平均エネルギ演算部５は１秒間の平均エネルギを求め、１秒単位で無音区間と有音区間の判定を行う。第３に有音区間であると判定された区間に対して、エネルギ変化率演算部８にて１秒毎のエネルギ変化率を求め、音声区間を抽出し、音声区間保持部１０にそのタイムコードを入力する。第４に、音声区間と判定されなかった区間に対して、平均バンドエネルギ比演算部１１にて１秒毎の平均バンドエネルギ比を求め、音楽区間を抽出し、音楽区間保持部１３にそのタイムコードを入力する。



【特許請求の範囲】

【請求項１】　オーディオ情報から音声区間と音楽区間を分類するオーディオ情報分類装置において、

入力されたオーディオ情報から単位時間ごとの周波数データを抽出するオーディオ周波数データ抽出手段と、

抽出した単位時間ごとの周波数データを用いて、その区間が無音区間か有音区間かを判定し有音区間のみを抽出する無音／有音判定手段と、

有音区間と判定された区間が音声であるか否かを判定する音声区間抽出手段と、　有音と判定された区間が音楽であるか否かを判定する音楽区間抽出手段とを具備したことを特徴とするオーディオ情報分類装置。

【請求項２】　請求項１のオーディオ情報分類装置において、

前記オーディオ周波数データ抽出手段によって抽出される単位時間ごとの周波数データは、入力されたオーディオ情報がＭＰＥＧ符号化データである場合、単位時間分のＭＰＥＧ符号化データの各フレームの先頭にあるサブバンドデータであることを特徴とするオーディオ情報分類装置。

【請求項３】　請求項１のオーディオ情報分類装置において、

前記無音／有音判定手段は、前記オーディオ周波数データ抽出手段により抽出された単位時間分の周波数データを用いて単位時間の平均エネルギを求め、該平均エネルギの大きさにより無音／有音区間を判定することを特徴とするオーディオ情報分類装置。

【請求項４】　請求項３のオーディオ情報分類装置において、

前記無音／有音判定手段は、入力されたオーディオ情報がＭＰＥＧ符号化データである場合、単位時間の平均エネルギは、ＭＰＥＧ符号化データの各フレームのサブバンドデータから求めたエネルギの単位時間における総和であることを特徴とするオーディオ情報分類装置。

【請求項５】　請求項１のオーディオ情報分類装置において、

前記音声区間抽出手段は、前記オーディオ周波数データ抽出手段により抽出した単位時間ごとの周波数データからエネルギ変化率を求め、該エネルギ変化率の大きさにより、音声区間を抽出することを特徴とするオーディオ情報分類装置。

【請求項６】　請求項５のオーディオ情報分類装置において、

前記音声区間抽出手段は、入力されたオーディオ情報がＭＰＥＧ符号化データである場合、エネルギ変化率は、ＭＰＥＧ符号化データのサブバンドデータから求めた隣り合うフレームの２つのエネルギの比の単位時間における総和であることを特徴とするオーディオ情報分類装置。

【請求項７】　請求項１のオーディオ情報分類装置において、

前記音楽区間抽出手段は、前記オーディオ周波数データ抽出手段により抽出した単位時間ごとの周波数データから平均バンドエネルギ比を求め、該平均バンドエネルギ比から音楽区間を抽出することを特徴とするオーディオ情報分類装置。

【請求項８】　請求項７のオーディオ情報分類装置において、

前記音楽区間抽出手段は、入力されたオーディオ情報がＭＰＥＧ符号化データである場合、平均バンドエネルギ比は、ＭＰＥＧ符号化データのサブバンドデータの全データに対する低周波帯域のサブバンドデータの割合であることを特徴とするオーディオ情報分類装置。

【発明の詳細な説明】

【０００１】

【発明の属する技術分野】本発明はオーディオ情報の分類装置に関し、特に、符号化されていない元のままのオーディオ情報あるいは符号化されたオーディオ情報から音声区間と音楽区間を分類できるオーディオ情報の分類装置に関する。

【０００２】

【従来の技術】オーディオに関する研究は、今までは、周波数解析等を用いて計算機上に取り込まれた音声が何であるかを認識する音声認識や、調音パラメータ等によって機械的に音声を合成する音声合成の分野に関する研究が主流であり、オーディオをタイトルや内容によって分類するインデクシングに関する研究はまだ少ない。インデクシングに関する研究報告としては、例えば、南、阿久津らの“大量映像に対する効率的アクセスインターフェース”、ITE Technical Report Vol.19,No.7,pp.1－６のように音楽区間を検出し、その情報を用いて音楽が流れている動画像をインデクシングするものがある。

【０００３】

【発明が解決しようとする課題】しかしながら、この研究報告は、音声区間の検出に関しては何ら触れておらず、音声区間を検出することができないという問題がある。また、該研究報告は、音楽区間の検出に関しても、周波数スペクトルのピークをＬＰＣケプストラムによって検出し、その平均持続時間を用いて音楽区間を検出しているため、圧縮符号化されたオーディオ情報からの検出は不可能であるという問題がある。

【０００４】本発明の目的は、前記した従来技術の問題点に鑑み、音楽区間のみならず音声区間の検出もできるオーディオ情報分類装置を提供することにある。他の目的は、符号化されていないオーディオ情報および圧縮符号化されたオーディオ符号化データの両方でオーディオ情報を音楽区間と音声区間に分類することを可能にするオーディオ情報分類装置を提供することを目的とする。

【０００５】

【課題を解決するための手段】前記目的を達成するため

に、本発明は、オーディオ情報から音声区間と音楽区間を分類するオーディオ情報分類装置において、入力されたオーディオ情報から単位時間ごとの周波数データを抽出するオーディオ周波数データ抽出手段と、抽出した単位時間ごとの周波数データを用いて、その区間が無音区間か有音区間かを判定し有音区間のみを抽出する無音／有音判定手段と、有音区間と判定された区間が音声であるか否かを判定する音声区間抽出手段と、有音と判定された区間が音楽であるか否かを判定する音楽区間抽出手段とを具備した点に第１の特徴がある。また、本発明は、入力されたオーディオ情報がＭＰＥＧ符号化データであっても、符号化データ上でオーディオ情報を音声区間と音楽区間に分類できるようにした点に第２の特徴がある。

【０００６】本発明によれば、符号化されていない元のままのオーディオ情報、あるいは符号化されたオーディオ情報のいずれからも、簡単かつ高速で、音声区間と音楽区間を分類することができるようになる。

【０００７】

【発明の実施の形態】以下に、図面を参照して、本発明を詳細に説明する。図１は本発明の一実施形態の構成を示すブロック図、図２、図３は、図１のシステムの動作、特に制御部２０の動作の概要を表わすフローチャートである。この実施形態は、動画像および音声符号化の国際標準方式であるＭＰＥＧ１（ＩＳＯ／ＩＥＣ　１１１７２）により圧縮された音声符号化データを用いて音声、音楽を分類するものであるが、本発明はこれに限定されるものではない。

【０００８】以下に、図１および図２、図３を参照して、本実施形態の構成と動作を説明する。図１に示されているように、圧縮符号化されたオーディオの符号化データａは、可変長復号部１に入力される。ここで、該圧縮符号化されたオーディオの符号化データの構造について、ＭＰＥＧ１を例にして図４を参照して説明する。ＭＰＥＧ１では、図示されているように、元のオーディオ信号ｐからサンプリングした５１２個のＰＣＭサンプルＰをサブバンド符号化して３２個のサブバンドデータＰｉ(n)（ｎ＝０，１，…，３１）を作り、それを時間的にサンプルをずらしながら３６回繰り返して、合計１１５２個のサブバンドデータを１フレームの符号化データＱとしている。

【０００９】前記した構造の符号化データＱが前記可変長復号部１に連続して入力してくると、該可変長復号部１はこれを各フレームのサブバンドデータに復号し、サブバンドデータサブサンプリング部２に出力する。いま、ある単位時間を１秒とすると、該１秒は図５のａのように３８フレームから構成されているので、可変長復号部１は１秒分の符号化データに対し、同図のｂのように、３８個の３２×３６サンプルを出力する。

【００１０】サブバンドデータサブサンプリング部２では、図５のｃに示されているように、単位時間（例えば、１秒）分のサブバンドデータのうち各フレームｉの先頭にあるサブバンドデータS_i(n)（ｉ＝０，１，…，j-1)を抽出し、図１の各フレームのエネルギ演算部３および第１のメモリ４に入力する。

【００１１】以上の動作は、図２では、ステップＳ１～Ｓ９で行われる。ステップＳ１では、フレーム番号を表すｉが０と置かれ、ステップＳ２ではサブバンド番号を表すｎが０と置かれる。ステップＳ３では、可変長復号部１にて符号化データが可変長復号され、ステップＳ４ではｉフレーム目の先頭のサブバンドデータS_i(n)が抽出される。次に、ステップＳ５にて、ｎ＝３２が成立するか否かの判断がなされ、この判断が否定の時にはステップＳ６に進んでｎに１が加算される。そして、ステップＳ３に戻って前記と同様の処理が行われる。以上のステップＳ３～Ｓ６の処理が繰り返し行われて、ステップＳ５の判断が肯定になると、ｉフレーム目の先頭のサブバンドデータS_i(n)が抽出されたことになる。

【００１２】ステップＳ５の判断が肯定になると、ステップＳ７に進み、ｉに１が加算される。次にステップＳ８に進み、ｉ＝ｊが成立するか否かの判断がなされる。この判断が否定の時にはステップＳ２に戻り、再びｎ＝０とされて、再度前記した処理が続行される。以上の処理が繰り返し行われ、ステップＳ８の判断が肯定になると、ｉ＝０～（ｊ－１）フレームの先頭のサブバンドデータS_i(n)が抽出されたことになり、ステップＳ９にて、これらのサブバンドデータS_i(n)は図１の各フレームのエネルギ演算部３および第１のメモリ４に転送されることになる。

【００１３】各フレームのエネルギ演算部３では、下記の(1)式に従って各フレームのエネルギＥｉを計算し、平均エネルギ演算部５および第２のメモリ６に入力する。

【００１４】

【数１】

$$Ei = \sum_{n=1}^{32} (S_i(n))^2 \qquad \cdots\cdots (1)$$

S_iはサブバンドデータ、ｎはサブバンド番号である。

各フレームのエネルギＥｉが計算されると、該エネルギＥｉはステップＳ１０にて平均エネルギ演算部５および第２のメモリ６に転送される。平均エネルギ演算部５では、下記の(2)式に従って入力された各フレームのエネルギから単位時間間の平均エネルギＡＥを計算し第１の判定部７に入力する（ステップＳ１１）。

【００１５】

【数２】

5

$$AE=\sum_{i=1}^{j} Ei \qquad \cdots\cdots (2)$$

jは１秒間のフレーム数である。

第１の判定部７では、入力された単位時間間の音声情報が無音であるのか有音であるのかを、下記の(3) 式に従って判定し条件に合う場合には有音であると判定する（ステップＳ１２）。有音である場合には無音である場合に比べて単位時間間の平均エネルギＡＥは大きいから、下記の(3) 式が成立することになる。

【００１６】$AE>\alpha$ …(3)

ここに、αは予め定められた第１の閾値である。

【００１７】該第１の判定部７において、入力された単位時間間の音声情報が有音であると判断された場合には、第２のメモリ６より各フレームのエネルギ単位時間分を読み出してエネルギ変化率演算部８に入力し（図３のステップＳ１３）、下記の(4) 式に従ってエネルギ変化率Ｃを計算し、第２の判定部９に入力する。一方、無音であると判定された時には、以降の音声、音楽判定処理を終了し、ステップＳ１に戻る。下式のＣは、ＭＰＥＧ符号化データのサブバンドデータから求めた隣り合うフレームの２つのエネルギの比の単位時間における総和を表している。

【００１８】

【数３】

$$C=\sum_{i=1}^{j} \left| 10\log_{10}(Ei+1/Ei) \right| \qquad \cdots\cdots (4)$$

音声の時間波形を見ると、単語や音節ごとに波形も変化し、その間は数１０ｍ秒にわたって無音となるため、そのスペクトル変化率は、連続波形となる音楽に比べて非常に大きくなる。そこで、第２の判定部９では入力された単位時間の音声情報が音声区間であるか否かを下記の(5) 式に従って判定し、条件に合う場合には音声区間と判定し、その区間のタイムコードｄを音声区間保持部１０に出力する（ステップＳ１４の判断が肯定、ステップＳ１５）。

【００１９】$C>\beta$ …(5)

ここに、βは第２の閾値である。

【００２０】一方、音声区間でないと判断された場合には（ステップＳ１４の判断が否定）、第１のメモリ４より各フレームの先頭のサブバンドデータを読み出して平均エネルギ比演算部１１に入力する（ステップＳ１６）。

【００２１】平均バンドエネルギ比演算部１１では、下記の(6) 式に従って平均バンドエネルギ比Ｂｍｉを計算して第３の判定部１２に入力する。

【００２２】

【数４】

6

$$Bmi=\frac{\sum_{n=0}^{k-1}(S_i(n))^2}{\sum_{n=k}^{31}(S_i(n))^2} \qquad \cdots\cdots (6)$$

音声の周波数は、図７(a) に示されているように、一般的に低周波帯域に集中し、一方音楽の周波数は、同図(b) に示されているように、全帯域に分散する傾向がある。換言すれば、音声のサブバンドデータが低周波帯域に集中するのに対して、音楽のサブバンドデータは全帯域にわたって分散する傾向がある。そこで、第３の判定部１２では、入力された単位時間間の音声情報が音楽区間であるか否かを下記の(7) 式に従って判定し（ステップＳ１７）、条件に合う場合には音楽区間と判定し、その区間のタイムコードｅを音楽区間保持部１３に出力する（ステップＳ１８）。

$Bmi<\gamma$ …(7)

ここに、γは第３の閾値である。

【００２３】以上のように、本実施形態によれば、圧縮符号化されたオーディオの符号化データから、音声区間と音楽区間を区別し、それぞれの区間のタイムコードを音声区間保持部１０および音楽区間保持部１３のそれぞれに記憶させることができるようになる。

【００２４】本発明は、さらに圧縮符号化されていないオーディオ情報の分類に関しても適応できる。その場合の実施形態を以下に示す。

【００２５】圧縮符号化されていないオーディオ情報を扱う場合は、図１の可変長復号部１およびサブバンドデータサブサンプリング部２は高速フーリェ変換部（以下、ＦＦＴ変換部と呼ぶ）に置き換えられる。元のオーディオ情報からこのＦＦＴ変換部において、図６にあるようにＦＦＴ変換を行い、単位時間分の周波数データを抽出する。今、該単位時間を１秒とすると、元のオーディオ信号ｐからサンプリングした２０４８個のサンプルをＦＦＴ変換し、それを時間的にサンプルをずらしながら３８回繰り返して、合計２０４８×３８個のＦＦＴデータを単位時間分の周波数データとしている。

【００２６】その後、各フレームのエネルギ演算部、平均エネルギ演算部、エネルギ変化率演算部、および平均バンドエネルギ比演算部で、それぞれ下記の(8) 式、前記(2) 式、(4) 式、および下記の(9) 式に従ってそれぞれ各フレームのエネルギＥｉ、平均エネルギＡＥ、エネルギ変化率Ｃ、平均バンドエネルギ比Ｂｍｉを計算し、第１の判定部７、第２の判定部９、第３の判定部１２にてそれぞれ無音／有音の判定、音声の判定、音楽の判定を行う。

【００２７】

【数５】

$$Ei = \sum_{n=1}^{2048} (F_i(n))^2 \quad \cdots\cdots (8)$$

$$Bmi = \frac{\sum_{n=1}^{m-1} (F_i(n))^2}{\sum_{n=m}^{2048} (F_i(n))^2} \quad \cdots\cdots (9)$$

【００２８】

【発明の効果】以上説明したように、本発明によれば、圧縮符号化された音声データから符号化データ上でオーディオ情報を音声区間と音楽区間に分類することが可能になるという効果がある。

【００２９】なお、本発明を実際に動作させたところ、次のような結果が得られた。すなわち、ＭＰＥＧ１レイヤ２で符号化された合計９０分のニュース番組と音楽番組のオーディオビットストリームを用いて、１秒毎の音声区間と音楽区間の分類を行った。音声区間は背景に音楽などがなく音声のみが含まれる区間を対象とし、音楽区間は音声の有無にかかわらず楽器演奏がある区間を対象とした。音声区間の検出に関しては、８９．４%、音楽区間に関しては７９．３%の検出率を得ることができ、音声区間の検出に関しては実用レベルの検出率を得ることができた。また、本発明によれば、圧縮符号化されていないオーディオ情報の分類に関しても、簡単に、音声区間と音楽区間に分類することが可能になるという効果がある。

【図面の簡単な説明】

【図１】　本発明の一実施形態のオーディオ情報分類装置の構成を示すブロック図である。

【図２】　図１の制御装置の動作を説明するためのフローチャートである。

【図３】　図２の続きの動作を説明するためのフローチャートである。

【図４】　ＭＰＥＧオーディオ符号化データの構造を説明するための図である。

【図５】　図１のサブバンドデータサブサンプリング部の動作を説明するための図である。

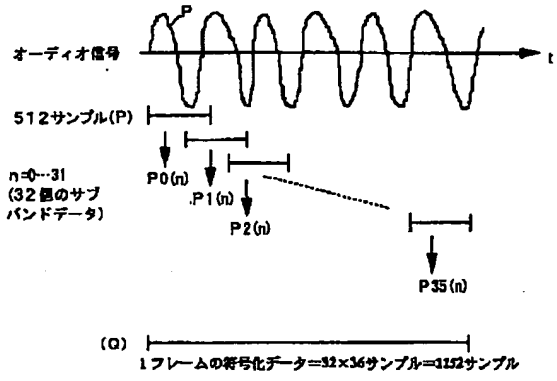
【図６】　符号化されていないオーディオ情報の周波数データの抽出方法を説明するための図である。

【図７】　音声と音楽の周波数分布の傾向を示す図である。

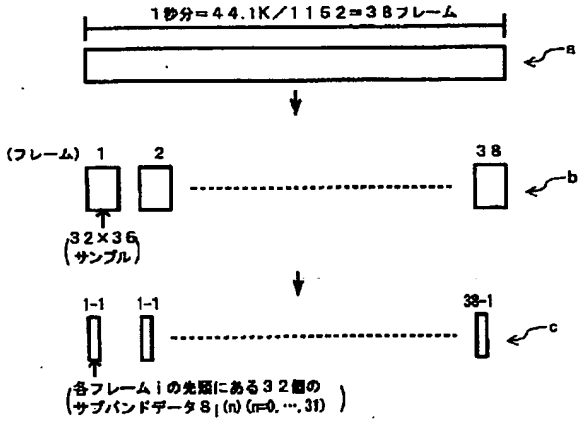
【符号の説明】

１…可変長復号部、２…サブバンドデータサブサンプリング部、３…各フレームのエネルギ演算部、４…第１のメモリ、５…平均エネルギ演算部、６…第２のメモリ、７…第１の判定部、８…エネルギ変化率演算部、９…第２の判定部、１０…音声区間保持部、１１…平均バンドエネルギ比演算部、１２…第３の判定部、１３…音楽区間保持部。

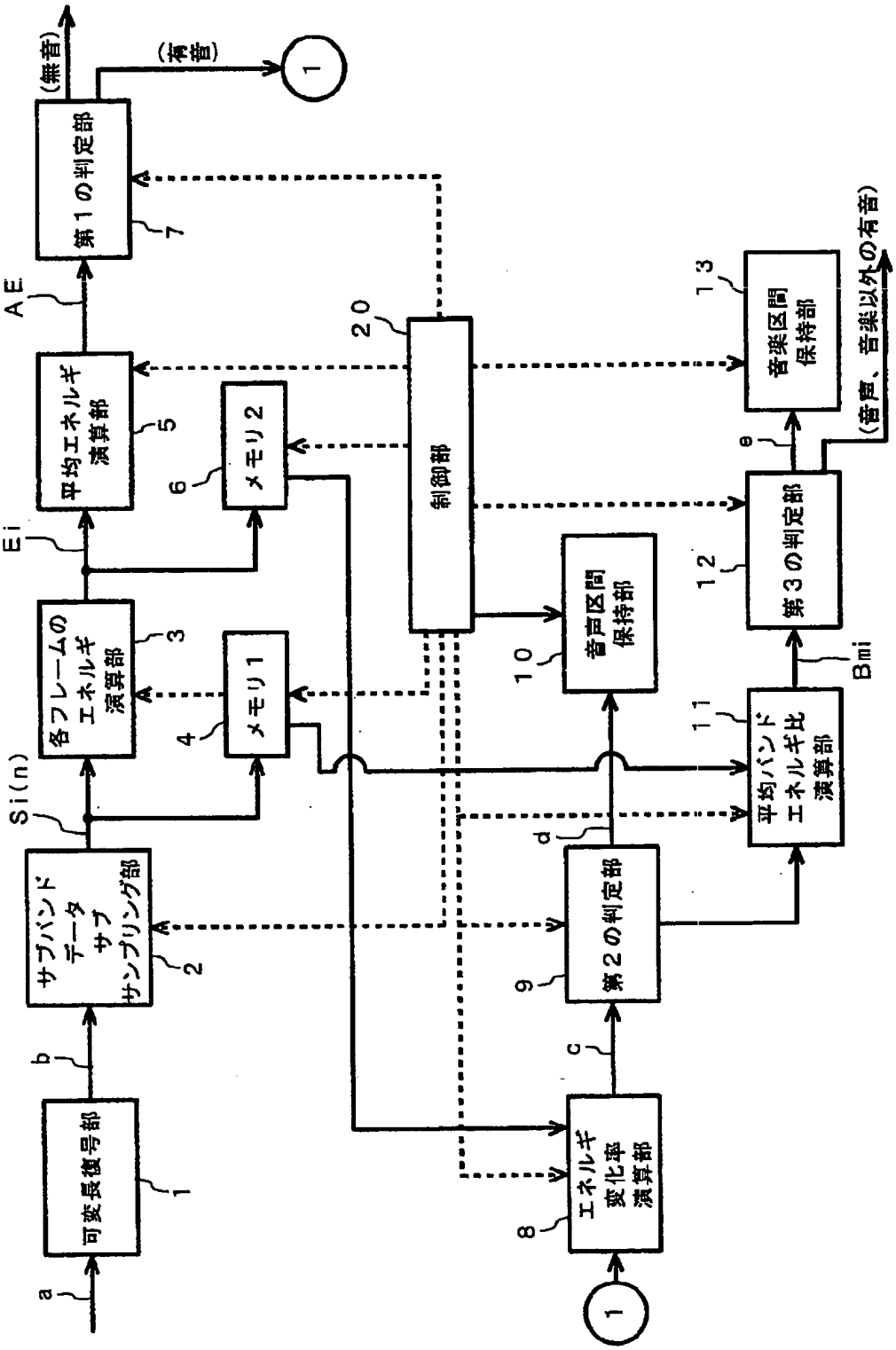
【図４】



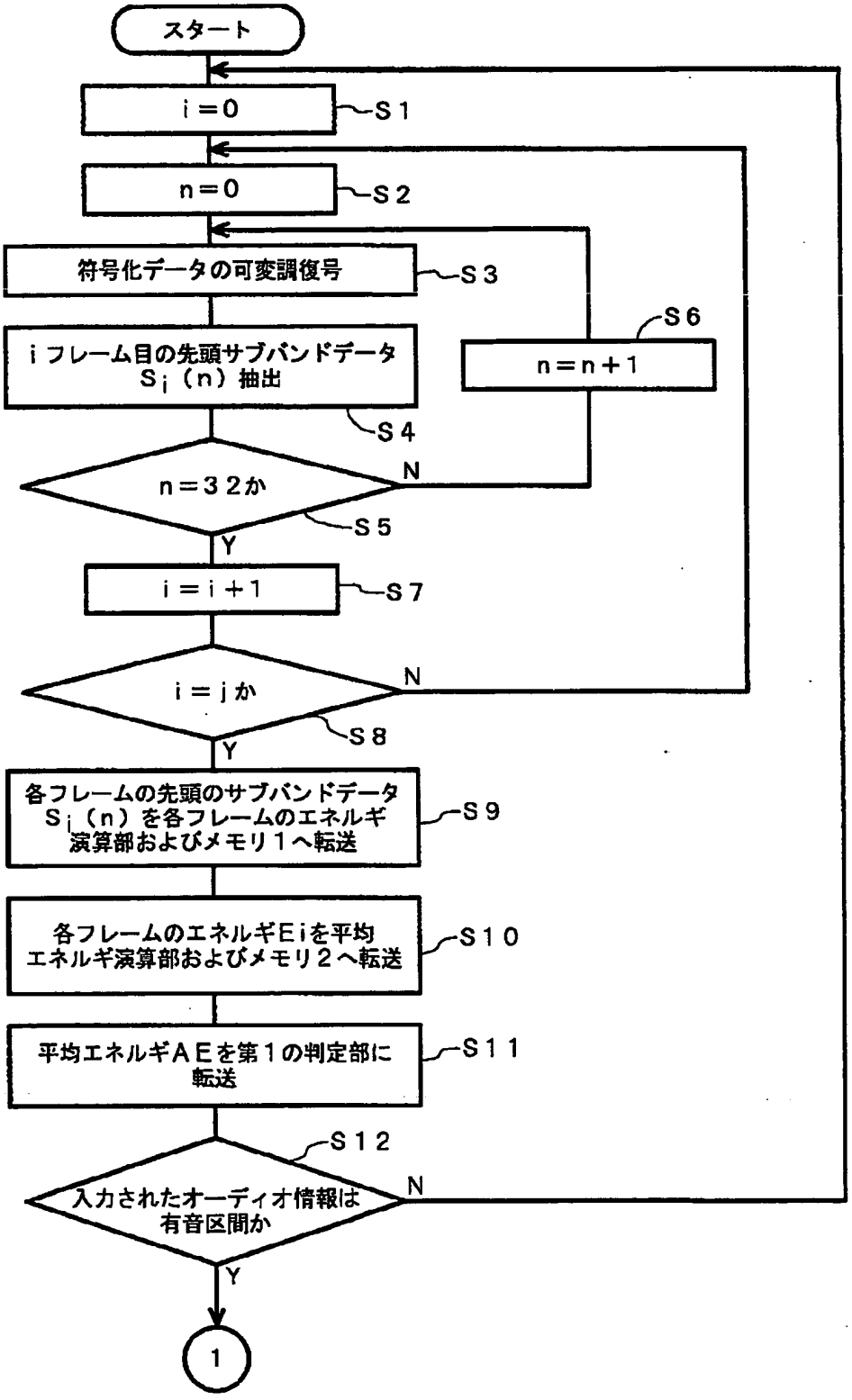
【図５】



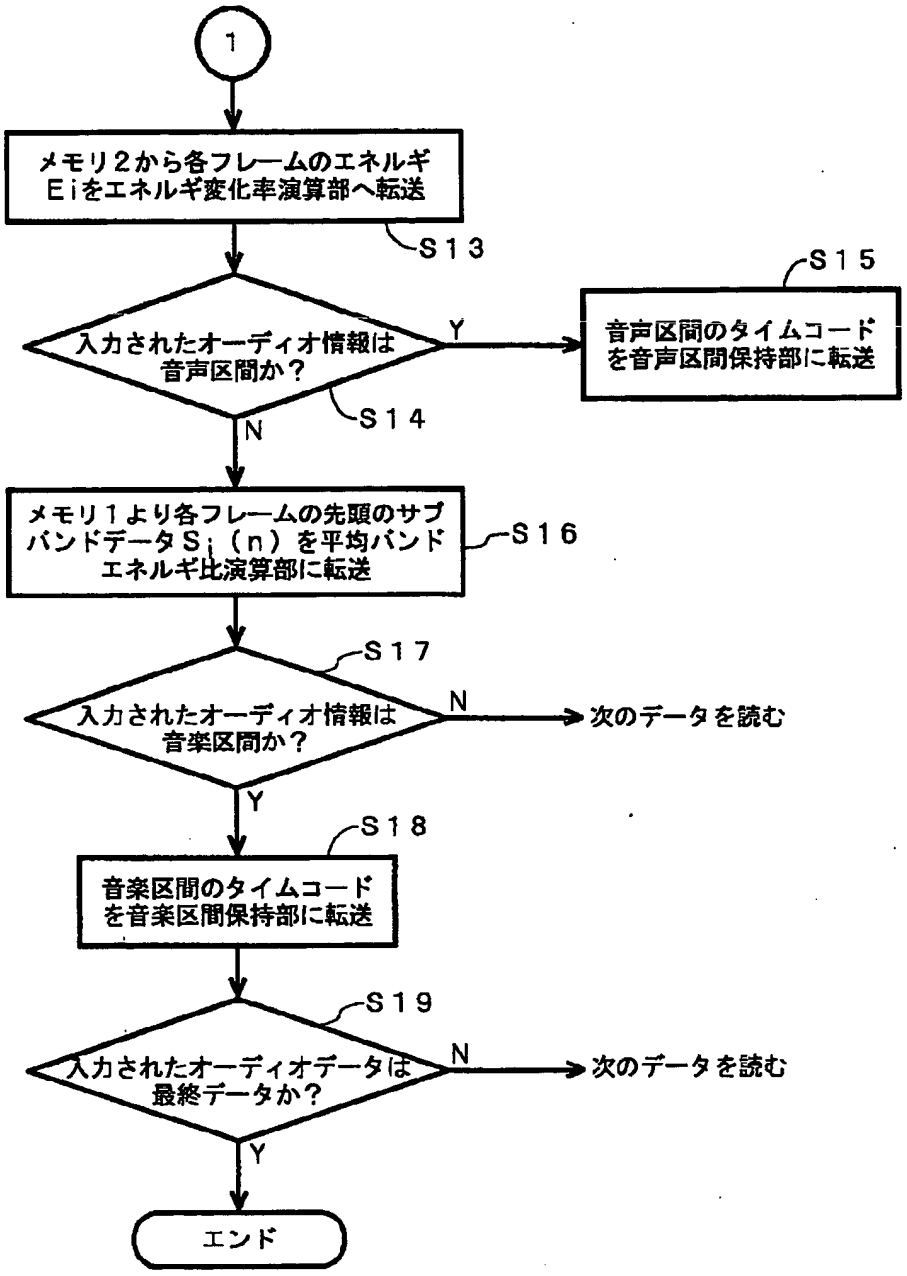
【図１】


可変長復号部
1
サブバンドデータサブサンプリング部
2
各フレームのエネルギ演算部
3
メモリ1
4
平均エネルギ演算部
5
メモリ2
6
第１の判定部
7
エネルギ変化率演算部
8
第２の判定部
9
音声区間保持部
10
平均バンドエネルギ比演算部
11
第３の判定部
12
音楽区間保持部
13
制御部
20
a
b
c
d
e
Si(n)
Ei
AE
Bmi
(無音)
(有音)
(音声、音楽以外の有音)
1
1

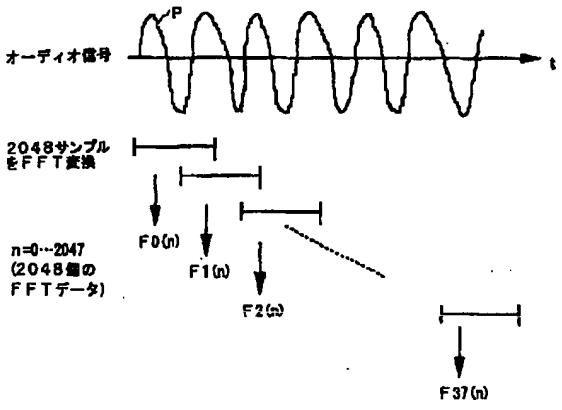
【図２】


スタート
i＝0
S１
n＝0
S２
符号化データの可変調復号
S３
iフレーム目の先頭サブバンドデータ
S_i（n）抽出
S４
n＝３２か
N
S５
Y
n＝n＋１
S６
i＝i＋１
S７
i＝jか
N
S８
Y
各フレームの先頭のサブバンドデータ
S_i（n）を各フレームのエネルギ
演算部およびメモリ１へ転送
S９
各フレームのエネルギEiを平均
エネルギ演算部およびメモリ２へ転送
S１０
平均エネルギＡＥを第１の判定部に
転送
S１１
S１２
入力されたオーディオ情報は
有音区間か
N
Y
1

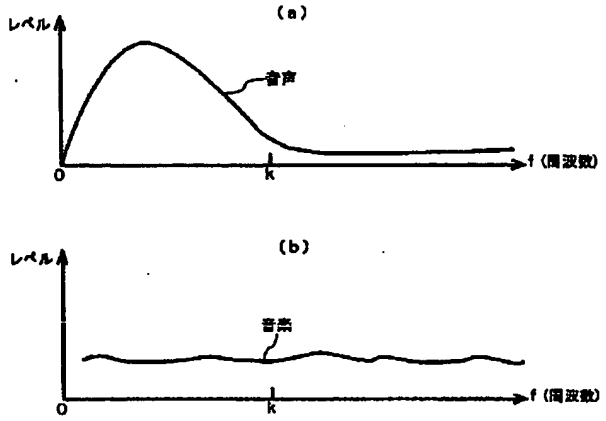
【図３】


1
メモリ２から各フレームのエネルギ
Ｅｉをエネルギ変化率演算部へ転送
Ｓ１３
入力されたオーディオ情報は
音声区間か？
Ｙ
Ｎ
Ｓ１４
Ｓ１５
音声区間のタイムコード
を音声区間保持部に転送
メモリ１より各フレームの先頭のサブ
バンドデータＳｉ（ｎ）を平均バンド
エネルギ比演算部に転送
Ｓ１６
Ｓ１７
入力されたオーディオ情報は
音楽区間か？
Ｎ
次のデータを読む
Ｙ
Ｓ１８
音楽区間のタイムコード
を音楽区間保持部に転送
Ｓ１９
入力されたオーディオデータは
最終データか？
Ｎ
次のデータを読む
Ｙ
エンド

【図６】


オーディオ信号
P
t
2048サンプル
をＦＦＴ変換
n=0…2047
(2048個の
ＦＦＴデータ)
F0(n)
F1(n)
F2(n)
F37(n)

【図７】


(a)
レベル
音声
0
k
f (周波数)
(b)
レベル
音声
0
k
f (周波数)